SINGER DIARIZATION: APPLICATION TO ETHNOMUSICOLOGICAL RECORDINGS

Marwa Thlithi
IRIT, Univ. Toulouse
118, route de Narbonne
31062 Toulouse -France
thlithi@irit.fr

Claude Barras
LIMSI-CNRS, Univ.
Paris-Sud, 91403,
Orsay, France
Claude.Barras@limsi.fr

Julien Pinquier, Thomas Pellegrini
IRIT, Univ. Toulouse
118, route de Narbonne
31062 Toulouse -France
{pinquier,pellegrini}@irit.fr

1. INTRODUCTION

A music audio document can be structured automatically by many ways according to the final objective. In the context of a project on indexing ethno-musicological audio documents, we asked ourselves the questions: who is singing and when. By analogy with *speaker diarization* which consists in detecting who is speaking and when, we called the fact of detecting changes of singers, *singer diarization*. Figure 1 illustrates the task. The ground truth consists of a manual annotation in singing turns, and eventual entry/exit of instruments.

In the context of the ANR DIADEMS[1] project *(Description, Indexing, Access to ethno-musicological and Sound Documents)* on indexing ethno-musicological audio documents, singer diarization automatically appeared to be essential. In this paper, we present our developed singer diarization system which is applied on ethno-musicological recordings.

The paper is organized as follows. In the next section, we describe our singer diarization system. In section 3, the application context is presented. In section 4, performance is presented and discussed.

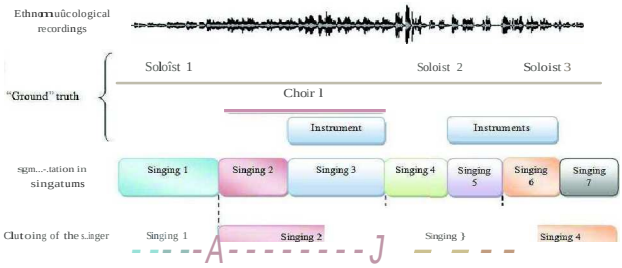

Figure 1. Illustration of singer diarization task.

2. SINGER DIARIZATION METHOD

Singer diarization consists in segmenting musical recordings, and then in labeling segments known as "acoustically homogeneous", our final goal is to obtain segments comprised of the singing of a single group of singers. Singer diarization is divided into two steps: segmentation step and clustering step.

Segmentation step consists in segmenting musical recordings into segments "acoustically homogeneous". Our method for the segmentation step is based on the Bayesian Information Criterion (BIC), which is widely used in audio segmentation (Chen, 1998; Delacourt, 2000; Siu, 1991). Its application on ethno-musicological recordings required an adaptation of two parameters: the size of the signal window, in which a border of segment is searched, and the penalty factor. The adaptation of the window analysis size was solved by implementing a version of the algorithm in which the window size increases while no potential boundary is found (Cettolo, 2005). For the penalty parameter, we observed that no single value was optimal for all the recordings. This led us to propose the Consolidated *A posteriori* Decision (DCAP) method, which consists in combining several segmentations obtained by varying the value of this parameter within the interval [0.8 1.2] with a step of 0.01 (Thlithi, 2014). Then, a vote is carried out on the candidates obtained from all these 41 segmentations: a boundary is validated if it was found by at least S_0 segmentations among all the segmentations. S_0 is determined on a development set. A tolerance gap of 0.5 s was used for this purpose. We used Mel Frequency Cepstral Coefficients (MFCC) for the parameterization step.

In order to achieve the goal of labeling all segments produced by the same group of singers with a unique identifier, an agglomerative clustering is performed on the output of the segmentation step. The signal is also parameterized with MFCC for this clustering step and clusters are modeled by a single Gaussian with a full covariance matrix. In the fust step, each segment seeds one cluster. The two nearest clusters according to the BIC criterion are then merged until the criterion sign changes.

3. CORPUS

The *"DIADEMS"* corpus was provided by the ethnomusicologist partners of the DIADEMS project. Examples are accessible online[2]. It is comprised of music recordings with a variable sound quality (outdoors in general, presence of background noise and audio events other than music). These records were done in several sub-Saharian countries. They mainly contain singer turns solo / choir with instruments or speech. This corpus contains 9 music recordings of 20 minutes in total which we divided into a development corpus (DEY) and an evaluation corpus (EVAL) in the proportions 20% and 80%.

[1] http://www.irit.fr/recherches/SAMOVA/DIADEMS/enlwelcome/

[2] http://diadems.telemeta.org/archives/fonds/CNRSMH_DIADEMS/

This corpus was manually annotated in terms of singer tums. A segment boundary is inserted in the following situations:

- Change from a group of N singers G; (i=l..N) to another group ofN' singers Gj' G=l..N'), such as:

  G; *f.* Gj' and 'V i, **j**
- Change from a group of singers to a no-singing area (silence, instruments, speech, etc.) and vice versa.

Then, the segments which contain the same group of singers are annotated with the same label.

4. RESULTS

We used the DEY subset to determine the S_0 parameter of the DCAP method for the segmentation step. We obtained S_0 equal to 15. A 61.2% F-m.easure was obtained for the segmentation step for the EVAL corpus.

Global results of singer diarization system on the DEY and EVAL corpus are presented in Table 1. Performance of the clustering result is expressed in term of diarization error rate (DER) which is the standard for speaker diarization and measures the fraction of time that is not attributed to the correct label, given an optimum mapping between the reference identifiers and the tru.e labels. The value of the penalty factor for clustering set on the DEV is equal to 6.

Corpus	DER
DEY	17.8%
EVAL	43.1%

Table 1. Global performance for DEY and EVAL corpus.

We noticed that performance varies significantly from one recording to another. Indeed, ethno-musicological recordings are very heterogeneous. Listening to the recordings where we have many errors reveals the presence of superimposed singers, rapid altemations between soloists and a choir, the presence of percussive instruments such as bells, hand elaps, and background noise. Moreover, these recordings proved to be more difficult to manually annotate in general.

5. CONCLUSION

In this article, we presented the singer diarization task. The long-term objective is indexing the content of ethnomusicological recordings. For the segmentation step, we applied a method based on the BIC criterion. The choice of optimal single value for the penalty parameter of this criterion proved unsatisfactory. In order to avoid selectmg a single value, we combined obtained segmentations with different values, and the final segmentation is obtained by keeping only the bmmdaries present in several of them. For the clustering step, we applied also the BIC criterion with an adjustment of its penalty parameter on the DEV set.

In order to improve our system, other clustering approaches are currently being tested such as the method of the ILP clustering with i-vectors.

6. ACKNOWLEDGEMENTS

This work was partly funded by the French National Agency for Research (ANR) under grant ANR-12-CORD- 0022-05 (project DIADEMS).

7. REFERENCES

Cettolo, M., Vescovi, M., & Ri2zi, R. (2005). Evaluation of BIC-based algorithms for audio segmentation. In *Computer SpeechAndLanguage,* (pp.l47-170).

Chen, S.S, & Gopalakrishnan, P.S, (1998). Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Infonnation Criterion. *The DARPA Broadcast News Transcription and Understanding Worlrshop.*

Delacourt, P., & Wellekens, C., (2000). DISTBIC: a speaker-based segmentation for audio data indexing. In *Speech Communication,* vol. 32, (pp.111-126).

Siu, M.-H., Yu, G., & Gish, H. (1991). Segregation of speakers for speech recognition and speaker identification. In *Proceeding of International Conference on Acoustics, Speech. and Signal Processing,* Toronto, Canada, (pp. 873-876).

Thlithi, M., Pellegrini, T., Pinquier, J., & An.dré-Obrecht, R., (2014). Segmentation in singer turns with the Bayesian Information Criterion. In *Proceedings of International Speech Cmnmunication Association,* Singapore, (pp. 1988-1992).